Data Collection and Preprocessing Phase

Date	25 September 2024
Team ID	SWTID1726888137
Project Title	intelligent handwritten digit identification system for computer applications
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification

For an intelligent handwritten digit identification system, establishing a robust data collection plan and identifying reliable raw data sources are foundational steps. This ensures that the system is trained on high-quality data, leading to accurate and generalized predictions across various use cases.

Data Collection Plan

Section	Description
Project Overview	This project aims to develop an intelligent handwritten digit identification system that leverages machine learning algorithms to accurately recognize and classify handwritten digits. The primary objective is to create a robust model that can be used in various applications, such as digit recognition for forms, checks, and digital handwriting analysis.
Data Collection Plan	Data will be collected from several reliable sources, including open-source datasets and user-generated inputs. The main sources are the MNIST dataset for initial training and custom data collected through user submissions via a web/mobile application to enrich the dataset with diverse handwriting styles.
Raw Data Sources Identified	1. **MNIST Dataset**: A widely used dataset containing 70,000 grayscale images (60,000 training and 10,000 testing) of handwritten digits. It serves as a benchmark for evaluating machine learning models. 2. **User-Generated Data**: A custom dataset collected from users through an interactive application that allows them to draw digits.

	This source captures a variety of handwriting styles and real-world digit formations. 3. **Synthetic Data Generation**: Utilization of data augmentation techniques (e.g., rotation, shifting, scaling) to create additional training samples from existing datasets, enhancing model robustness against variations in input.

Raw Data Sources

Source Name	Description	Location/URL	Format	Size	Access Permissions
MNIST Dataset	A dataset of handwritten digits containing 70,000 grayscale images (60,000 training and 10,000 testing) commonly used for training and evaluating machine learning models.	MNIST Dataset	CSV	0.5 GB	Public